La anonimización: un instrumento clave para una gestión de datos eficiente

La anonimización de conjuntos de datos requiere una metodología de alto nivel para encontrar soluciones al problema de compartir información sin violar el derecho a la privacidad de las personas. La cuestión requiere regulaciones basadas en una visión técnica y política.

El uso de datos es una herramienta imprescindible a la hora de diseñar y ejecutar políticas públicas. Hoy en día, los organismos públicos tienen varios incentivos para no compartir datos, sobre todo por los riesgos —incluso penales— que pueden correr si esos datos son reidentificados. Pero la solución de no compartir es la peor estrategia, porque se pierde la riqueza de la utilización de datos, que es cruzar información de diferentes fuentes: **la anonimización es una metodología de trabajo recomendable porque reduce sensiblemente estos riesgos.**

MENSAJES CLAVE

- Buscar alternativas y establecer buenas prácticas para compartir los datos con un nivel de detalle suficiente como para utilizar algoritmos avanzados es clave a la hora de poner en valor la información.
  - **Anonimizar** es eliminar la posibilidad de asociar registros de datos con los individuos a los que esos registros se refieren.
  - **Desidentificar** es eliminar elementos que asocien un registro a una persona individual, como códigos de identificación personal, códigos de dispositivos (direcciones IP, MAC) e identificadores biométricos.
- Hay una tendencia, sobre todo en la Administración Pública, a armar **"silos de datos"**: grupos separados y autónomos que disponen de información y la usan pero no la comparten, por lo que quedan aislados entre sí.
- La anonimización puede aportar seguridad a funcionarios y agentes públicos para que puedan intercambiar información entre organismos y potenciar el uso de datos en la toma de decisiones.
- Los datos pueden ser considerados **sensibles** por muchos motivos: origen racial y étnico, opiniones políticas, patrimonio personal, convicciones religiosas, afiliación sindical o información referente a salud o vida sexual. Todo dato es sensible si causa un daño a la persona en el momento en que se hace público.

Para seguir leyendo

Este informe de políticas públicas se desprende del documento **Anónimos pero no tanto: cómo hacer una gestión de datos eficiente sin poner en riesgo la privacidad**, elaborado por el Área de Datos de Fundar.

Cuatro etapas para una metodología eficiente en la gestión de datos

1. Identificar datos

Individualizar los campos que pueden actuar como identificadores y que deben ser borrados, modificados o eliminados: códigos, códigos de aparatos (direcciones IP, MAC) e identificadores biométricos. En esta etapa, es necesario **identificar qué propiedad de los datos se quiere conservar.** Entender para qué hacen falta los datos y cuál es su valor para encontrar soluciones adecuadas.

2. Identificar riesgos

Analizar los riesgos potenciales de la publicación, cruce y reidentificación de una base de datos es importante a la hora de entender los escenarios de uso. En muchos casos, el riesgo surge al cruzar los datos con otras bases. El análisis no debe hacerse solo de la base que se pretende publicar, sino de esa base en el contexto de su uso. **La identificación de riesgos no es absoluta ni objetiva: está asociada a una preocupación de un grupo de ciudadanos o stakeholders.**

3. Identificar soluciones

- **Privacidad diferencial:** una técnica que permite que cada consulta realizada a una base de datos no se responda con información exacta, sino que se introduzca cierta cantidad de ruido para evitar que se pueda usar una estrategia de reidentificación.
- **K-Anonimización:** no siempre se puede anonimizar completamente al individuo, pero si las pistas nunca nos llevan a grupos más pequeños que K individuos (es decir, si nunca distingo un individuo de un grupo de menos de K), se dice que existe K-anonimización. Por ejemplo, si la única información que brindo sobre una persona es que se recibió en 2003 en un colegio, y las personas que se recibieron ese año en el mismo colegio, fueron 200, ese dato está 200-anomizado, ya que el individuo queda "agrupado": es imposible diferenciarlo de los otros 200 para los cuales ese dato coincide. Por lo tanto, se dice que es 200-anónimo.
- **Eliminación:** siempre es posible eliminar una o más columnas o variables de una base de datos.
- **Generalización:** agrupar datos, brindar información solo a nivel de grupo (distribución estadística y parámetros generales), cambiar su nivel de detalle (es clave mantener la granularidad de las observaciones y, si es estrictamente necesario, hacer una agregación sobre la variable considerada sensible, pero solo sobre esa). Lo que no es una estrategia inteligente es solo "agregar" los datos (es decir, publicar un promedio o la suma de todos ellos) en forma descuidada como única opción.
- ***Hashing* y encriptado:** aplicar a un dato una función "de un solo sentido", es decir, una función cuya inversa es muy costosa de calcular.
- **Distorsionar los datos** ("agregar ruido"), mezclar, confundir. Esto suele ser usado en gráficos, pero se puede aplicar en varios contextos.

4. Identificar ataques y problemas

Establecer un criterio adversarial: se debe considerar que se enfrenta a un adversario y revisar la solución desde el punto de vista de un posible ataque. Esta visión es dinámica: aplicar una solución o una técnica seguramente genere un cambio de estrategia en el potencial ataque.